ANALES CERVANTINOS, VOL. LI,
PP. 231-250, 2019, ISSN: 0569-9878, e-ISSN: 1988-8325
https://doi.org/10.3989/anacervantinos.2019.011

# Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*

JUAN CEREZO SOLER*
JOSÉ CALVO TELLO**

**Resumen**
La estilometría se ha consagrado en los últimos años como uno de los métodos de investigación más sólidos en el campo de las Humanidades Digitales. Su eficacia como método para la investigación de autorías en textos anónimos está probada, tanto en casos de autores españoles como extranjeros. En las siguientes páginas se propone, justamente, la aplicación de esta herramienta digital para el esclarecimiento de la autoría de *La conquista de Jerusalén*, atribuida desde su descubrimiento a Miguel de Cervantes.

**Palabras clave:** estilometría; Humanidades Digitales; teatro; Miguel de Cervantes.

**Title: Authorship and style. A Cervantes attribution from the digital humanities. The case of *La conquista de Jerusalén***

**Abstract**
Stylometry has become, in recent years, one of the most solid research methods in the academic field of Digital Humanities. Its effectiveness as a method for authorship attribution has been proven with cases in Spanish and other languages. This paper applies one of these digital methods for the clarification about the hypothesis of *La conquista de Jerusalén*, attributed to Miguel de Cervantes since its discovery.

**Keywords:** Stylometry; Digital Humanities; Theatre; Miguel de Cervantes.

* Universidad Autónoma de Madrid. *juan.cerezosoler@gmail.com* / ORCID iD: https://orcid.org/0000-0002-1780-7973.
** Universidad de Würzburg. *jose.calvo@uni-wuerzburg.de* / ORCID iD: https://orcid.org/0000-0002-1129-5604.

**Cómo citar este artículo / Citation**
Cerezo Soler, Juan y José Calvo Tello (2019). «Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*», *Anales Cervantinos*. 51, pp. 231-250, https://doi.org/10.3989/anacervantinos.2019.011.

## 1. Introducción

Van ya veinticinco años del descubrimiento de *La conquista de Jerusalén* entre los fondos manuscritos de la Real Biblioteca de Palacio[1]. No escapará al lector interesado que han sido años de intensa investigación en los que todas y cada una de las aportaciones sobre la comedia han ido dirigidas a consolidar –o desmentir– la primera hipótesis de autoría que lanzó, en su momento, su descubridor, el profesor Stefano Arata. Hoy, prácticamente toda la comunidad investigadora acepta la paternidad cervantina de la obra ante el peso de los numerosos argumentos que la respaldan, si bien se muestra aún cauta a la hora de hablar de pruebas definitivas[2]. Efectivamente, en tanto no se descubra un testimonio textual que vincule de forma inapelable el nombre de Cervantes con el de la comedia palatina, no podremos aceptar su autoría más que como hipótesis. Altamente probable, sí, pero hipótesis al fin y al cabo.

Sin embargo, y al margen de la prudencia del investigador a la hora de formular este tipo de hipótesis, o de su elocuencia a la hora de respaldarlas, es forzoso declarar que los estudios de atribución de autoría han adolecido, hasta hace bien poco, de una falta de protocolo o metodología concreta con la que cimentar sobre la objetividad científica la validez de las investigaciones. Y el caso de *La conquista de Jerusalén*, en este sentido, no es una excepción. Así lo reclamaban Javier Blasco y Cristina Ruiz Urbón en 2009, a propósito de la extensa anonimia que encontramos en los textos españoles del Siglo de Oro:

> Aunque los estudios de atribución son tan viejos como la literatura misma, solo desde hace algunas décadas se ha realizado un esfuerzo digno de elogio por incorporar las fórmulas y los procedimientos que, desde la década de los 60, se han ido consolidando en el terreno de la Lingüística

1. Para un acercamiento a los fondos teatrales de esta biblioteca, *cf.* Stefano Arata (1989; 1991; 1992) y Rojo Alique (1996-1998).

2. Puede encontrarse un detallado estado de la cuestión en el capítulo dedicado a *La conquista de Jerusalén,* en la última edición del teatro completo de Miguel de Cervantes, bajo el sello de la Real Academia de la Lengua y al cuidado general de Luis Gómez Canseco. Fausta Antonucci, que es quien firma el capítulo (2015: II, 186-194), ofrece un minucioso comentario al problema de la autoría acompañado de un aparato bibliográfico actualizado hasta 2015, que es el año de la publicación.

> forense [...], apuesta interesante de cara a la construcción de una metodología fiable (por su capacidad de objetivación de los fenómenos observados) y al establecimiento de unos protocolos de actuación en los que se redujese notablemente el espacio concedido a la subjetividad o a la elección arbitraria por parte del analista (2009: 28).

Reivindicaban los citados investigadores la fijación de una metodología que no deje espacio a la intuición, la subjetividad, la querencia o la parcialidad del crítico[3] a la hora de lanzar una posible autoría sobre un texto anónimo. Y la hallaban felizmente en el análisis del discurso, del *usus scribendi*, genuino, inconsciente e indefectiblemente propio de cada autor. En definitiva: en el análisis del estilo. En esta línea se incluyen las presentes páginas, dedicadas a respaldar, de nuevo, la autoría cervantina de *La conquista de Jerusalén* a través de un método sólido, cerrado y homogéneo, basado exclusivamente en el análisis cuantitativo textual y contrastado con las últimas investigaciones sobre estilometría, tanto para la literatura española como extranjera.

## 2. El procedimiento estilométrico para el esclarecimiento de autorías

La estilometría se ha consagrado en los últimos años como uno de los métodos de investigación más sólidos en el campo de las Humanidades Digitales. Consiste, como bien señala el rótulo, en el análisis cuantitativo (*metría*) de los rasgos de *estilo* para la investigación literaria. Se ha usado, principalmente, para el rastreo de autorías, sí, pero también es herramienta eficaz a la hora de esclarecer matices tales como el género literario, la época de redacción (*estilocronometría*) o, incluso, el género del autor[4].

El impacto y la relevancia de la estilometría no se limita exclusivamente a los círculos de investigación y experimentación de las Humanidades Digitales. Muchos de los resultados estilométricos son capaces de incidir o modificar el fenómeno de recepción de obras y autores llegando, en algunos casos, a condicionar estrategias comerciales en el campo editorial o alterando la recepción de los textos más importantes de diversas tradiciones y culturas, tal es el caso, por ejemplo, de los escritos fundacionales, folklóricos o de tipo claramente religioso. Uno de los primeros trabajos sobre estilometría, publicado por Mosteller y Wallace hace ya más de cincuenta años (1963), se ocupó de los llamados *Federalist papers*, una colección de artículos periodísticos

3. Esta intención está presente en buena parte de los investigadores que, a día de hoy, se adentran en el estudio de atribuciones de autoría. El mismo equipo, un año más tarde, coordinaría la publicación de un volumen con valiosísimas aportaciones al asunto que nos ocupa (Blasco, Marín Cepeda y Ruiz Urbón 2010).

4. Como señala Oakes (2009: 1071), en traducción nuestra, la estilometría se fundamenta sobre la simple idea de que «aunque los autores puedan modificar de manera consciente su estilo, siempre habrá un uso inconsciente y consistente de rasgos estilísticos en sus obras».

publicados a finales del siglo XVIII con propaganda favorable a la ratificación de la constitución estadounidense. Todos los artículos fueron publicados por tres autores diferentes (Hamilton, Jay y Madison) y la autoría de doce de ellos estaba, en aquel momento, en discusión. El trabajo, pionero en cuanto al tipo de métodos que hoy proponemos, señaló con poco margen de duda que Madison había sido el autor de todos aquellos textos dubitados.

Un caso de investigación estilométrica que alcanzó gran repercusión en los medios fue el de Patrick Juola (2015) sobre la novela *The Cuckoo's Calling*, traducida como *El canto del cuco* y publicada bajo el seudónimo de Robert Galbraith. Patrick Juola, impulsado por el diario *The Sunday Times*, abordó desde la estilometría el rumor de que tras el seudónimo con el que se firmaba la novela se encontraba, en realidad, J. K. Rowling, la autora de la saga *Harry Potter*. Juola comparó la nueva novela con otra de corte policiaco escrita por Rowling, tratando ambas como entradas de un corpus más amplio. Los resultados señalaron una gran cantidad de rasgos estilísticos significativamente similares a los de Rowling. El periódico, con este aval estilométrico, contactó con la editorial de la autora, que terminó reconociendo su identidad tras el seudónimo de Galbraith.

En 2016 se publicó *The New Oxford Shakespeare: The Complete Works*, unas obras completas de William Shakespeare editadas por Taylor, Jowett, Bourus y Egan. Esta nueva edición contó con la novedad de un consejo sobre autoría (*Attribution Board*), formado por un equipo encabezado por Craig, Egan y Gants y especializado tanto en materia de atribuciones como, específicamente, en el método estilométrico. No escapa a nadie que los problemas de autoría sobre la obra de Shakespeare son, efectivamente, numerosos y complejos. Ello propicia que la investigación estilométrica no solo ocupe el contenido íntegro de uno de los volúmenes de esta publicación (volumen *Authorship Companion*), sino que haya servido como base para la definición y fijación del canon del autor.

Y estos son solo algunos ejemplos. En el ámbito literario español, muy por el contrario, no deja de sorprender la escasa cantidad de investigaciones dedicadas a abordar la autoría desde un planteamiento estilométrico. Siendo, además, el Siglo de Oro un espacio en el que los textos con autoría problemática constituyen más la norma que la excepción. Uno de los trabajos más completos de los publicados en los últimos años es el realizado por De la Rosa y Suárez (2016: 373–438) sobre la autoría del *Lazarillo*. Del análisis de un amplio corpus configurado con la nómina de autores que, según la crítica tradicional, pudieron haber escrito el *Lazarillo*, se arrojaron resultados bastante dispares: Juan de Arce y Otálora, Alfonso de Valdés y, en menor medida, Cristóbal de Villalón y Pedro Mejía serían los candidatos más cercanos al estilo del *Lazarillo*. La variabilidad de estos resultados no resta, creemos, solidez al método estilométrico, sino que señala con elocuencia la enorme complejidad del caso concreto del *Lazarillo*: la existencia de numerosísimos candidatos muy dispares entre sí, la escasa disponibilidad de muchas de sus obras y una falta de homogeneidad genérica tal que puede, llegado el caso,

contaminar los resultados estilométricos. Sobre esto hay que añadir la posibilidad, siempre presente, de que el verdadero autor haya quedado, por el motivo que fuera, excluido del corpus analizado, lo que inutilizaría cualquier investigación estilométrica desarrollada sobre el mismo.

Igualmente complejo es el caso abordado por Rißler-Pipka (2016) sobre la autoría del *Quijote* apócrifo. Los resultados señalaron que el autor estilométricamente más cercano al tal Avellaneda es nada menos que Cervantes. Una interpretación estilométrica tradicional de estos resultados nos llevaría a señalar al complutense como el autor escondido tras la firma de Avellaneda. La sola propuesta es peregrina, pues tal interpretación pasaría por alto las insalvables coincidencias que rodean a los tres *Quijotes*: género, época, personajes, temas, localizaciones, etc. Al eliminar del corpus los dos *Quijotes* de Cervantes, los resultados se mostraron variables, y todo parecía apuntar a que ninguno de los candidatos recogidos era, en realidad, Avellaneda. El mismo problema ha sido revisado recientemente (Blasco 2016: 97-115) en un estudio en el que se perfiló la configuración del corpus de candidatos, aportando con ello datos muy significativos sobre la atribución autoral en cada una de las secciones del libro, en coherencia, además, con lo intuido por buena parte de los críticos especializados en la figura de Avellaneda[5].

Se trata de dos casos enormemente atractivos para la crítica, dada la relevancia que tendrían los avances, en caso de darse, para la historia de la literatura. Pero no son los únicos. Otros investigadores han arrojado luz sobre problemas de autoría en casos de menor calado, tales como Herrera (Hernández Lorenzo 2017), Góngora (Rojas Castro 2017) o Lope de Vega[6].

Con todo, y al margen de los trabajos particulares ya mencionados, también es escasa la producción científica en español sobre el método estilométrico en sí. Calvo Tello, en 2016, halló la explicación de este silencio crítico en las siguientes razones: en primer lugar, en que la mayoría de los proyectos, iniciativas de documentación y, sobre todo, herramientas informáticas hayan sido desarrolladas por investigadores angloparlantes. Pero el motivo fundamental es que, a diferencia de las otras lenguas europeas, el español sigue sin contar con un repositorio de textos digitales en un formato óptimo para la investigación (XML-TEI).

Desde un punto de vista más técnico, las diferentes metodologías estilométricas varían en cuanto a los algoritmos utilizados y a los tipos de rasgos extraídos del texto. Tras décadas de afinación, pueden asentarse tres tipos de

5. No estará de más enfatizar aquí lo dicho por Blasco a propósito del corpus utilizado en el análisis estilométrico: «Para que el análisis estadístico y estilométrico ofrezca resultados con ciertas garantías de objetividad, hay que proceder con rigor en la constitución del corpus de análisis, en el tratamiento del mismo y, sobre todo, en la interpretación de los resultados» (2016: 103). Sobre estas huellas de, digámoslo así, obsesión por la objetividad, se sitúa nuestro estudio.

6. Tal y como han expuesto recientemente José Calvo Tello y Gaston Gilabert en el XI Congreso de la Asociación Internacional del Siglo de Oro, en la Universidad Complutense de Madrid, en julio de 2017.

señales: la frecuencia de los caracteres de puntuación[7], la presencia de elementos gramaticales sencillos –como la longitud de palabras u oraciones– y la frecuencia léxica (Stamatatos 2009: 3). Es este último signo el que ha proporcionado resultados más fiables en los últimos años. De hecho, en la actualidad, prácticamente toda la comunidad estilométrica acepta el rango de palabras más frecuentes como el parámetro de búsqueda más útil, centrando, por tanto, la discusión en problemas de ajuste del método: cuál es la cantidad óptima de palabras más frecuentes o *most frequent words* (MFW) –desde pocas decenas hasta varios miles– para que el resultado sea fiable, cómo tratar su coaparición –a veces frecuente– con otras palabras (*ngramas*) o su presencia en el conjunto del corpus (*culling*).

Como sea, puede afirmarse que el estudio estilométrico goza hoy de una larga andadura y que ha alcanzado la madurez suficiente como para respaldar con solvencia las hipótesis de autoría lanzadas sobre buena parte de los textos anónimos en nuestra literatura. Siempre, claro está, que se den las condiciones necesarias para el análisis, tales como la existencia de un corpus de candidatos lo más cerrado y completo posible y la disponibilidad de varios textos por autor con los que configurar una nómina abundante que permita, llegado el momento, extraer conclusiones fiables sobre la factura estilística, no solo de la obra que se pone en cuestión sino de todos y cada uno de los autores investigados. Esta es la postura desde la que abordaremos el análisis de *La conquista de Jerusalén,* sin más intención que la de aprovechar esta nueva herramienta digital para esclarecer aún más, si se puede, la hipótesis sobre su verdadero autor[8].

## 3. Análisis de *La conquista de Jerusalén*[9]

### *3.1. Método de aprendizaje automático*

La primera de las preocupaciones que debe asaltar al investigador, al hilo de lo expuesto más arriba, es la configuración del corpus textual. Para el caso de *La conquista de Jerusalén* hemos optado por considerar como posibles

7. Muy eficaz a la hora de identificar el comportamiento estilométrico de un autor pero inútil, en todo caso, para los textos del Siglo de Oro, pues la puntuación era, casi siempre, obra del cajista o editor.

8. Es bueno que insistamos, primero, en el carácter provisional de los resultados, pues por elocuentes que puedan parecer, no constituyen en sí mismos una prueba definitiva. Segundo, en que la objetividad deseada nos obliga a no tener un candidato preasignado, ya sea por querencia o convencimiento intelectual. Para el caso de *La conquista de Jerusalén* hemos tratado a todos los autores como posibles candidatos, preparándonos, en todo caso, a que los resultados arrojaran datos ajenos a la hipótesis cervantina. Solo así se puede, como se verá, lanzar la conclusión sobre autoría después de realizar el análisis, y no antes.

9. Los resultados que aquí mostramos vienen a completar, con nuevos y más sólidos datos, el primer acercamiento al tema llevado a cabo por Calvo Tello y Cerezo Soler (2018).

autores a todos los que conforman la generación teatral de 1580[10], aquel grupillo de dramaturgos que ocupó la escena española hasta la llegada, dramática para algunos, de Lope de Vega. Pero elaborar un corpus que pueda utilizarse es algo más complejo que la simple recopilación de nombres y obras. No estarán de más algunas nociones de cara a que el mismo método sea repetido en el futuro. Para empezar, hay que tener presente que el análisis estilométrico no solo da información sobre el autor de un texto. Esta es, según terminología especializada, la *señal más fuerte*, pero no es la única. El sexo del autor (Argamon *et al*. 2002), el género literario (Kestemont 2011), la época de composición (Jockers 2014), son solo algunos de los signos que pueden vislumbrarse tras las pruebas estilométricas. El corpus verdaderamente útil será, por tanto, el diseñado para anular, en la medida de lo posible, todas estas señales secundarias a nuestro objetivo principal, tanto en la elección de los textos como en el tratamiento estructural y ortográfico de los mismos.

En consecuencia, hemos recopilado un total de diecisiete textos teatrales, en verso, escritos por los siete autores −todos varones− que configuran la nómina canónica de este teatro prelopesco. Dejamos a un lado, por tanto, cualquier obra en prosa escrita por nuestros candidatos, ya que es el estilo teatral la principal pista sobre la que edificaremos, más adelante, el rastreo estilométrico. La selección, además, se ha llevado a cabo según los siguientes criterios: 1) cada autor estará representado, como máximo, por tres obras; 2) como la obra dubitada es una comedia, se dará prioridad a las obras pertenecientes a este subgénero teatral; 3) la datación es, también, rasgo influyente en el análisis, por lo que se intentará, en la medida de lo posible, incluir las obras más cercanas cronológicamente a *La conquista*[11].

Con todo, el siguiente obstáculo que ha de salvar el investigador es, ahora sí, el acceso a las obras en un formato que permita su tratamiento en el campo de las Humanidades Digitales. A la espera de un proyecto de investigación que posibilite la libre disposición de textos literarios en español, hemos realizado, para el caso que nos ocupa, una labor de extracción desde diversas fuentes. Notamos que cada uno de los textos a los que habíamos accedido se había tratado según los criterios de su editor, siguiendo, como es lógico, distintos procedimientos de modernización. Con el objetivo de anular cualquier

10. Así ha de ser, dadas las fechas de composición que se manejan sobre la comedia: una evidente cesura temática en el tramo final evidencia que la obra original se construyó sobre la estructura de cuatro actos, y no tres, como quiso la receta teatral lopesca (Arata 1992: 11; 1997: 56); al mismo tiempo, todos los especialistas dedicados al teatro del siglo XVI coinciden en que una obra como esta tuvo que redactarse en fechas inmediatamente posteriores a 1580 y en todo momento, previas a 1586 (Brioso Santos 2009: 26-30).

11. Estos han sido los criterios que han guiado la selección del corpus. Como es lógico, no siempre han podido cumplirse todos, pues en nuestra nómina hay autores de quienes solo se han conservado una o dos obras teatrales. También se ha tenido que decidir sobre si priorizar un criterio sobre otro. En el caso del mismo Cervantes, por ejemplo, se tuvo que decidir si incluir sus tragedias −más cercanas cronológicamente a *La conquista de Jerusalén*− o sus comedias −de redacción mucho más tardía−.

diferencia textual que pudiera contaminar los resultados, se ha optado por unificar según una modernización casi completa, respetando únicamente las formas contraídas *–dello, desto, aquesto–* para evitar la incisión sobre el comportamiento métrico de la obra; del mismo modo y por el mismo motivo, se ha mantenido la forma arcaica infinitivo+pronombre *–decilla, matallo–*. Por último, se ha eliminado del corpus todo lo que no fuera texto puro, es decir, encabezamientos, títulos, listas de personajes, acotaciones, nombres de interloctuores, etc. (como suele ser habitual en trabajos sobre teatro y como está implementado en *stylo* cuando se utiliza TEI para teatro); dejando para el análisis, simplemente, el contenido de los parlamentos teatrales como la materia netamente literaria, útil para la indagación del comportamiento estilístico del autor y a la que llamaremos, en adelante, *texto plano*.

Hecho y definido el corpus, toca analizarlo. Utilizaremos la herramienta *stylo* (Eder, Kestemont y Rybicki 2016) en sus opciones por defecto, salvo en los datos que especifiquemos más adelante, llamándolo desde *Python* mediante la librería *rpy2*. Los datos (en forma de tablas de frecuencias léxicas, como ya hiciese Fradejas, 2016) y los pasos exactos dados en la programación (versión de *stylo*, parámetros y valores utilizados, etcétera) pueden ser consultados en el *Jupyter Notebook* (también guardado como PDF) localizado en el repositorio *GitHub*[12]. De esta manera, compartimos con la comunidad investigadora los datos, los detalles del software e incluso un entorno en el que los resultados puedan ser verificados con facilidad. Las versiones de Delta probadas han sido: *Classic Delta* (también conocida como *Burrows Delta*), *Eder's Delta* y *Cosine Delta* (presentada por Smith y Aldridge en 2011 y ratificada por Jannidis *et al*. 2015). Los resultados de las tres versiones son muy similares, aunque la evaluación de *Eder* aporta mejores resultados[13]. Será la que incluyamos en este artículo, aunque en el repositorio pueden consultarse los resultados de las otras versiones.

En primer lugar, y antes de aplicar la herramienta sobre *La conquista de Jerusalén*, es necesario comprobar que la organización automática de los textos (*clustering*) coincida con la realidad. Ello corroborará nuestros criterios en la elaboración del corpus:

12. Accesible en: <https://github.com/morethanbooks/publications/tree/master/Cervantes_Conquista>. Fecha de acceso: 22 de noviembre de 2019.

13. *Cosine Delta* tiende a confundir los textos de la Cueva y de Lasso, mientras que *Eder's Delta* tiende a clusterizarlos y clasificarlos correctamente.

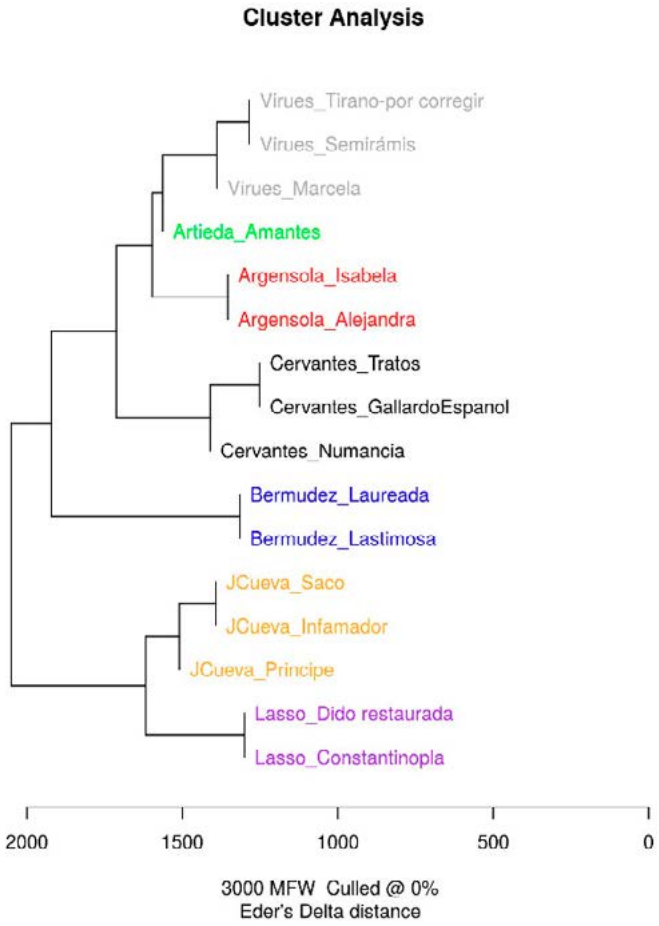


FIGURA 1. *Cluster* de textos de atribución segura.

Puede observarse que los textos de Virués, Argensola, Cervantes, Bermúdez, Juan de la Cueva y Lasso de la Vega aparecen correctamente distribuidos, por lo que se puede reconocer, sin mucho problema, a un único autor en cada una de las diferentes ramas o agrupaciones de obras. Bien, una vez confirmado que tanto el algoritmo como los parámetros seleccionados para el análisis funcionan, pues no ha mezclado ni confundido a ninguno de los candidatos, añadamos el texto discutido al corpus y observemos los resultados:

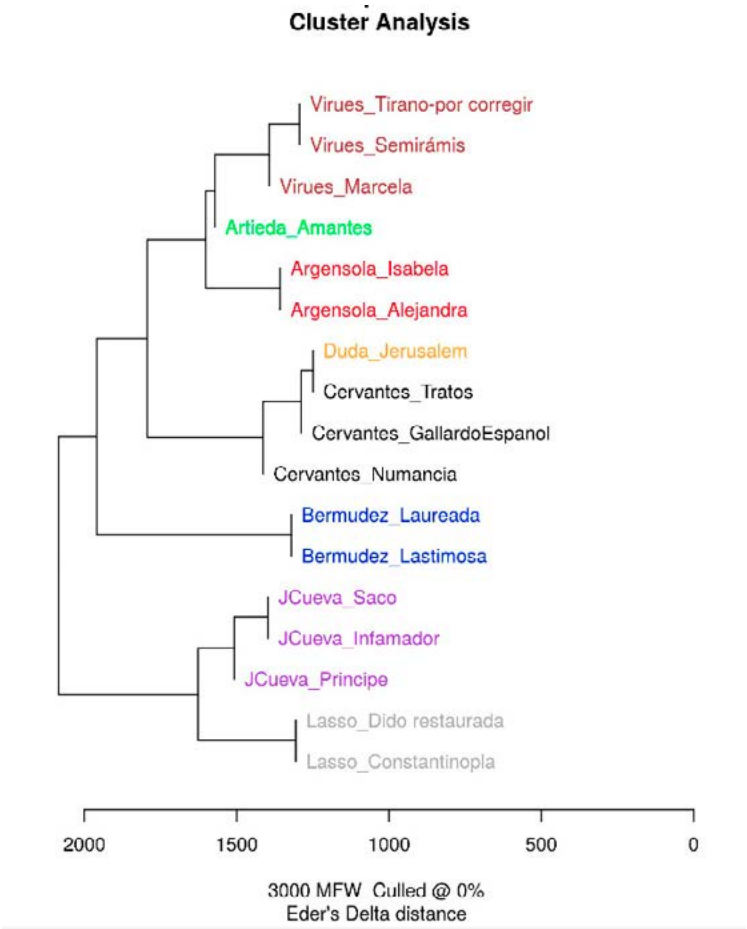


FIGURA 2. *Cluster* completo.

Como se observa, el programa ha agrupado el texto de *La conquista de Jerusalén* junto al resto de obras de Miguel de Cervantes. Esto, de momento, solo significa que el estilo literario de la comedia discutida es más similar al de las obras cervantinas que al del resto de obras recogidas en el corpus. Seamos prudentes. Solo se ha realizado un análisis sobre las 3000 palabras más frecuentes. Bien pudiera ser que al modificar este parámetro, los resultados varíen. Es posible –y necesario– cerrar el cerco estilométrico sobre *La conquista de Jerusalén*, para lo que habrá que afinar y ajustar, como decimos, los parámetros de rastreo. Así, si ampliamos el rango de búsqueda léxica y realizamos varios experimentos en serie, analizando desde las 500 hasta las 5000 palabras más frecuentes, podremos asentar sobre base sólida si la hipótesis cervantina es la correcta. Seleccionemos la cantidad que seleccionemos, se verá que en todas las pruebas el análisis estilométrico arroja como resulta-

do, efectivamente, a Miguel de Cervantes. Así lo declara este *consensus tree*, gráfico que condensa en una sola imagen los resultados de los últimos diez experimentos realizados:

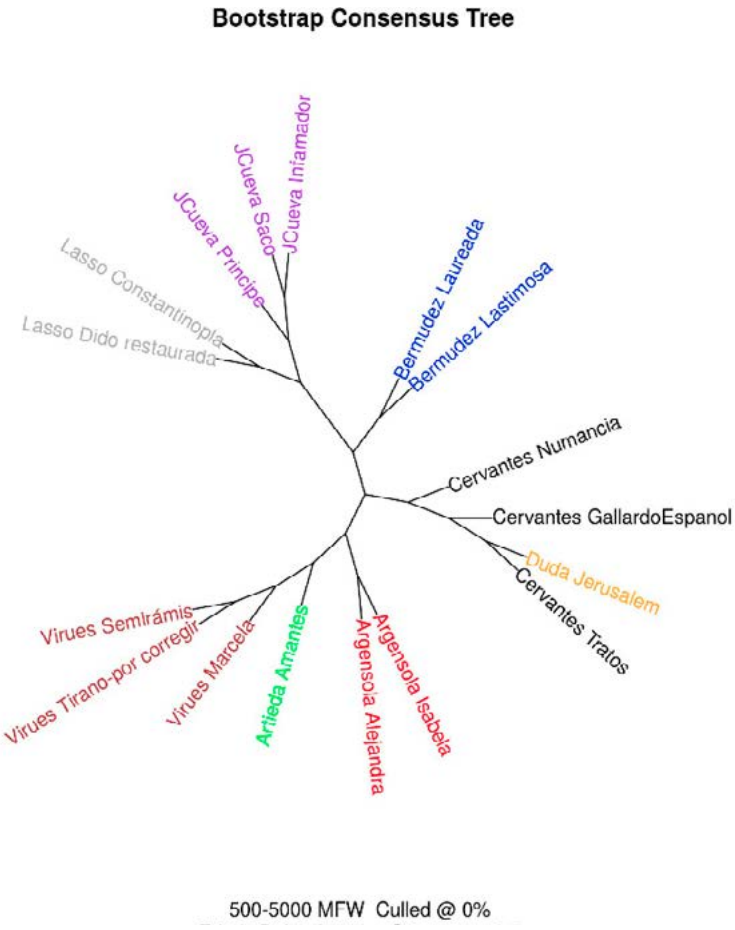


FIGURA 3. *Consensus tree*.

Bien se percibe que la inclusión de *La conquista de Jerusalén* en el grupo de textos cervantinos no se da exclusivamente tras la búsqueda de una cantidad específica de palabras más frecuentes, sino que es estable, manteniendo el vínculo con el autor del *Quijote* en todos y cada uno de los diferentes rangos aplicados.

Esta serie de resultados legitima desde la estilometría la hipótesis cervantina, principalmente frente a los autores representados en el corpus. Ello, quizá, permita ahora una nueva afinación del corpus dirigida a dificultar, un poco al menos, esta vinculación automática de *La conquista de Jerusalén* con los textos cervantinos.

No se nos escapa que los tres textos elegidos para representar a Cervantes –*Los tratos de Argel*, *La Numancia* y *El gallardo español*– comparten con el texto discutido una serie de características –subgénero, época de redacción, lugar de la acción, tema– que podrían comprometer el verdadero resultado estilométrico, es decir, que lo que podríamos estar observando no solamente sea la información de autor (Schöch 2013; Calvo Tello *et al*. 2017), sino el resto de semejanzas. Por ello, se ha optado por sustituir estos primeros textos de Cervantes por otros más lejanos, tanto en la cronología como en la factura, y mucho menos representativos de su quehacer teatral general. Son textos que, digámoslo así, podrían situarse en las antípodas de *La conquista de Jerusalén*, tal es la distancia que presentan con la comedia palatina en todos los aspectos. Esto nos permitirá forzar un poco más la maquinaria estilométrica para comprobar que, efectivamente, el análisis del estilo es capaz de trascender las semejanzas formales de los textos. Se ha seleccionado, pues, *La casa de los celos* y *Pedro de Urdemalas*. Aquí los resultados:

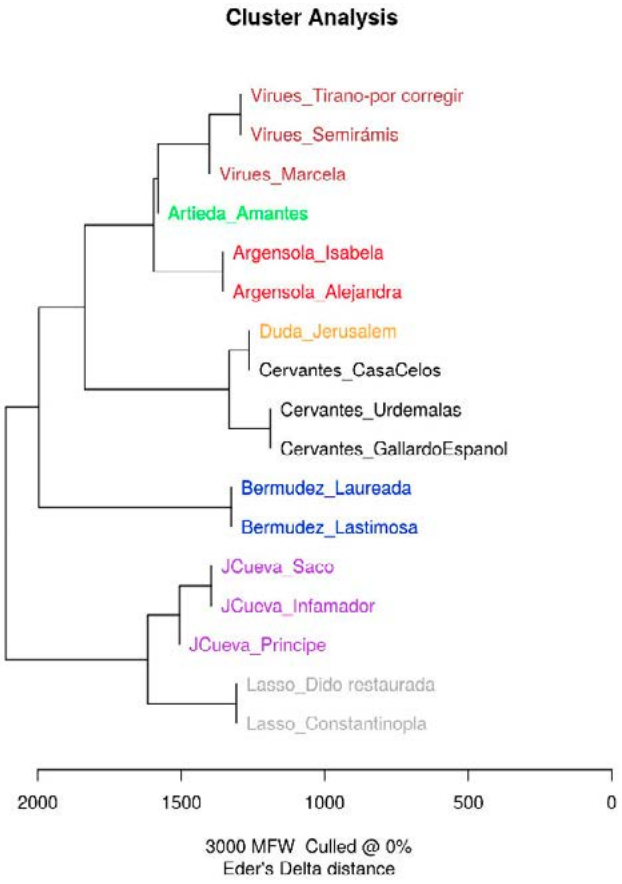


FIGURA 4. *Cluster* con textos menos representativos de Cervantes.

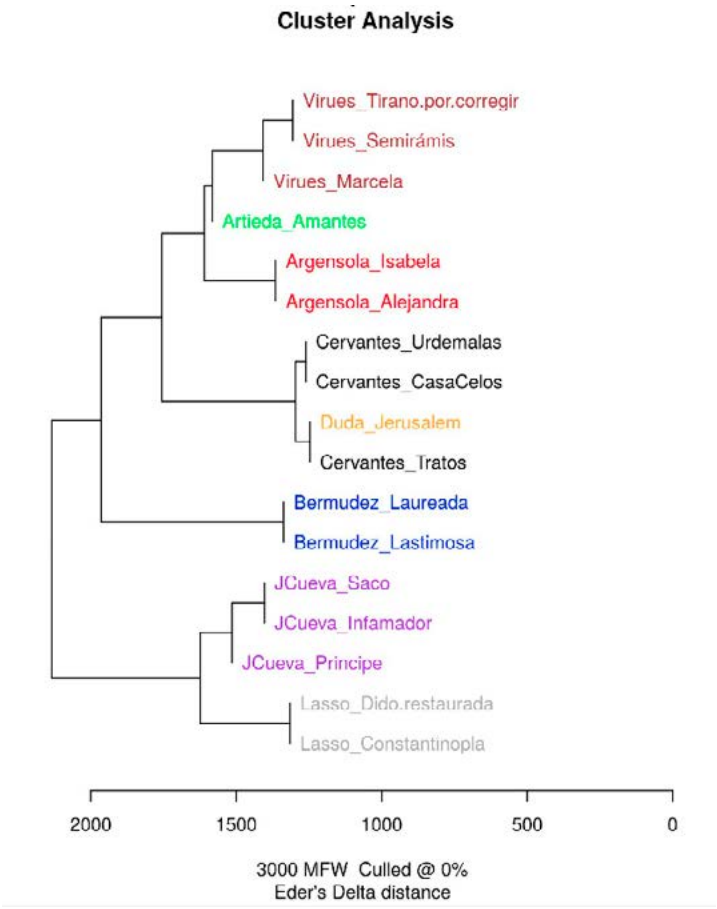


FIGURA 5. *Cluster* II con textos menos representativos de Cervantes.

Como se puede observar, el resultado es siempre el mismo. En primer lugar, la paternidad de Cervantes sobre estas dos nuevas obras es detectada correctamente. Esto significa que el algoritmo no necesita conocer al Cervantes más representativo para identificar su estilo frente al resto de autores. En segundo lugar, *La conquista de Jerusalén* es organizada, de nuevo, junto al resto de obras firmadas por el manco, dejando ya un escaso margen de duda sobre su autoría.

## *3.2. Clasificación*

Pero el experimento debe continuar. Hasta ahora hemos utilizado técnicas de agrupamiento automático, también llamado aprendizaje no supervisado.

Vamos a pasar ahora a lo que en informática se denomina *técnicas de aprendizaje supervisado*, muy útiles a la hora de realizar preguntas más específicas sobre el propio corpus y evaluar la precisión de los resultados.

El procedimiento es el siguiente: se empieza por la división del corpus en tres subcorpus: el de *aprendizaje*, el de *evaluación* y el de *test*. Los dos primeros sirven para que el algoritmo aprenda y reconozca los rasgos estilísticos de cada uno de los autores recogidos. Por ello, aquí se incluirá representación textual de todos ellos. Una vez "aprendidos" los rasgos textuales de cada autor, el algoritmo intentará una clasificación del subcorpus de *evaluación*; como esos textos no son discutidos, podremos evaluar la efectividad del algoritmo en función del porcentaje de acierto en la clasificación de obras por autores. El caso problemático, o texto discutido –aquí solo uno– se colocará en el subcorpus de *test*, y el algoritmo intentará asignarlo a alguna clase de las aprendidas, en este caso a algún autor. Evidentemente, la seguridad del resultado final irá indefectiblemente ligada al éxito de la evaluación, pues si la clasificación de los textos sobre los que no hay duda resulta errónea, todo resultado posterior sobre la obra en cuestión perderá credibilidad.

Y así las pruebas. Se han colocado los textos del corpus al completo en el subcorpus de *aprendizaje*. Para la *evaluación* hemos añadido los textos menos representativos de Cervantes: *La casa de los celos* y *Pedro de Urdemalas*. Finalmente, *La conquista de Jerusalén* como única entrada en el tercer subcorpus, el de *test*. Note el lector que este experimento no es automático, va dirigido y pautado en todo momento por el investigador, por lo que solo sirve para apuntalar una determinada hipótesis en base a los resultados obtenidos, concretamente, en la matriz *Delta* (*Eder's Delta* 3000 MFW):

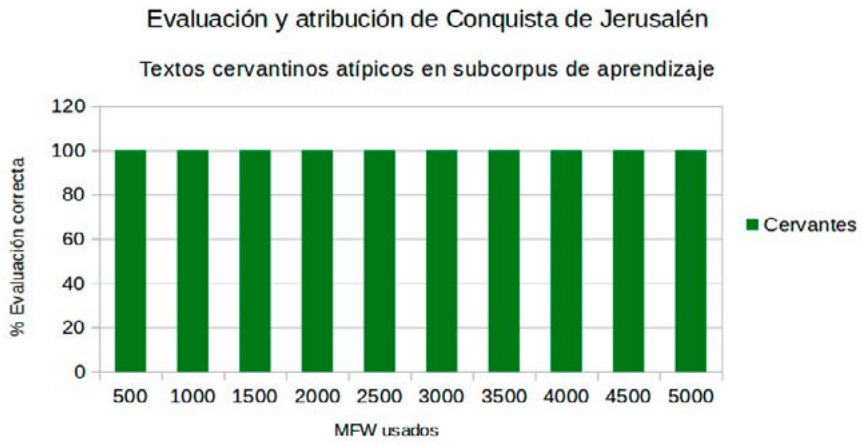


FIGURA 6. Resultados de atribución cervantina y *evaluación* usando los textos de *aprendizaje*.

Los resultados son absolutamente claros, dado el porcentaje de evaluación correcta que puede observarse: los textos cervantinos son vinculados siempre (desde las 500 hasta las 5000 palabras más frecuentes) a su legítimo autor. Y *La conquista* también.

En segundo lugar, y para finalizar, añadimos a este último corpus otros dos corpus colocando en el subcorpus de *aprendizaje* los textos que aquí hemos dado en llamar "menos representativos" de Cervantes junto con, al menos, otros dos textos del resto de autores. En el corpus de *evaluación* dejamos el resto de texto para poder cuantificar en qué medida el algoritmo clasifica correctamente los textos no discutidos. El objetivo en este caso es doble: por un lado, evaluar de nuevo si las atribuciones a otros autores son también correctas en diferentes posibles combinaciones y, por otro, dificultar la prueba forzando que el algoritmo aprenda los rasgos de Cervantes a través de sus textos menos representativos, y comprobar, además, si los más representativos mantienen su correcta clasificación. Los resultados se visualizan como diagrama de caja (cada diagrama de caja con tres datos de *evaluación*, uno por cada subcorpus) que muestran los resultados de la *evaluación*; en el eje horizontal, junto con la cantidad de MFW analizada, se señala el autor que el algoritmo seleccionó para el texto de *La Conquista de Jerusalén*.

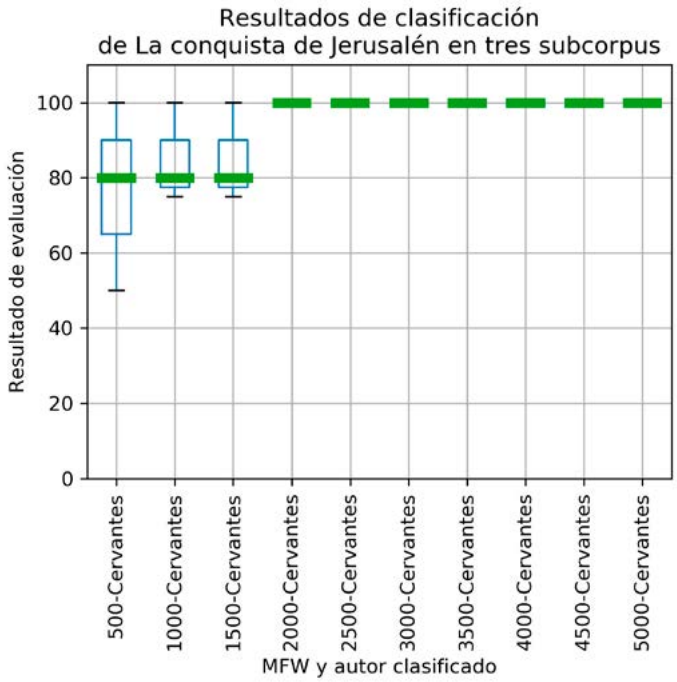


FIGURA 7. Resultados de atribución cervantina y de *evaluación* en tres subcorpus.

Como se observa, la evaluación fluctúa entre el 100% y el 50% mientras utilicemos 1500 MFW o menos. Una vez superamos la barrera de las 2000 MFW, el algoritmo clasifica correctamente todos los textos las tres ocasiones. Es decir, la evaluación de la clasificación no contiene errores: no tenemos razones para pensar que el algoritmo no esté consiguiendo clasificar los textos correctamente. En cuanto a los resultados del texto en discusión, el algoritmo en cada uno de los tres corpus, independientemente de la cantidad de rasgos, clasifica *La conquista de Jerusalén* bajo la autoría de Miguel de Cervantes.

## 4. Algunas conclusiones: el teatro del xvi, *La conquista de Jerusalén* y Miguel de Cervantes

Todos los resultados revelan que *La conquista de Jerusalén* tiene un lugar claro entre las obras de Miguel de Cervantes. La solidez del método estilométrico para las atribuciones de autoría viene refrendada, insistamos, no solo por el vínculo arrojado entre el autor deseado y la obra en cuestión, sino por la correcta distribución de todo el corpus teatral sometido a estudio. El mismo método que ha incluido la comedia anónima entre las de Cervantes también ha situado correctamente al resto de autores en sus respectivas ramas y con sus respectivas obras. Esto, al margen de la alta apariencia de fiabilidad que muestra en la asignación de autorías, proporciona al investigador una serie de claves para interpretar, siempre a la luz de los resultados, las relaciones entre los distintos autores de una misma generación.

Merced a estos resultados, y sin salir del corpus cervantino, podemos concluir que *La conquista de Jerusalén* presenta una relación de especial cercanía con *Los tratos de Argel* (Fig. 2). Esta cercanía bien puede entenderse en dos términos: cronológica y temática. Sendas comedias fueron compuestas en el mismo período de tiempo, concretamente en el que corresponde a su primera etapa de actividad teatral, alrededor de 1582, cuando Cervantes, recuperada ya su libertad, retomó su actividad literaria con la creación de varias piezas, muchas de ellas hoy perdidas. Sin embargo, mucho más relevante que esta cercanía cronológica resulta la relación temática entre ambas obras. Tanto *La conquista de Jerusalén* como *Los tratos de Argel* sitúan la trama teatral en un marco de conflicto con el enemigo musulmán. La primera, en la cruzada medieval sobre Tierra Santa capitaneada por Godofre de Bullón; la segunda, en pleno escenario de cautiverio norteafricano, en el siglo XVI. Pero las dos con un enemigo claramente identificado en la época como antagonista: el musulmán. Ambas obras se desarrollan, por así decir, en el contexto de un enfrentamiento de calado profundamente religioso, lo que se dejará notar en el registro léxico utilizado por Cervantes y, por lo tanto, en los resultados estilométricos.

Así ocurre, de hecho, no solo entre *La conquista* y *Los tratos*, como se ha visto; sino también con *El gallardo español*. El método estilométrico plasma

en los resultados las similitudes de tema y contenido como un parámetro de relación, en este caso, más sólido que las coincidencias de género, subgénero o cercanía cronológica. Al conectar –pese a tratarse de una composición tardía– *El gallardo español* con las dos obras antes mencionadas, se nos está revelando la tendencia de la estilometría a agrupar obras con contenido léxico similar, lo que, a su vez, es muy útil para clasificar la producción literaria de un único autor. Para el caso concreto de Cervantes, el vínculo entre *El gallardo* y *La conquista* o *Los tratos* encaja cómodamente con la poética del conflicto como clave interpretativa de toda la dramaturgia cervantina[14].

Un poco más alejada, aunque igualmente identificada con Cervantes, aparece en estos primeros análisis *El cerco de Numancia*. Dos motivos justifican, esta vez, el distanciamiento: argumento dramático y subgénero teatral. *La Numancia* es la obra de datación más temprana de todas las de Cervantes, muy cercana cronológicamente a *Los tratos de Argel* y, creemos, también a *La conquista de Jerusalén*. Y sin embargo, como apuntábamos más arriba, la relación de contenido prima sobre la cercanía o lejanía cronológica, al menos en términos puramente estilométricos. El argumento de *La Numancia*, aunque en líneas generales trate del asedio militar a una ciudad, se desarrolla no solo en un marco temporal muy lejano al siglo XVI, sino en un escenario que nada tiene que ver con el del resto de obras analizadas. En *La Numancia*, los personajes y tipos humanos que protagonizan la acción son totalmente ajenos al binomio cristianismo-islam que late en el resto de obras propuestas y que, además, son muy del gusto teatral cervantino. No aparece en ninguno de sus actos un solo conflicto en clave religiosa[15], lo que, de nuevo, ilustra este distanciamiento estilométrico.

Pero únicamente con esto bien se nos podría acusar de manipular los resultados, eligiendo las obras cervantinas más cercanas −tanto temática como cronológicamente− a *La conquista de Jerusalén*. Por ello, y siguiendo la pista de estos primeros resultados, se impuso la necesidad de adaptar el corpus, cambiando estas primeras obras de Cervantes por otras especialmente lejanas, tanto en género y contenido como en estilo y forma, a la misma *Conquista de Jerusalén*. Se trataba, en definitiva, de restarle al análisis la

14. Esta clave permite la agrupación de las obras según el tipo de conflicto llevado a escena. Así, tres obras presentarían el conflicto religioso del cautiverio (*El trato de Argel*, *Los baños de Argel* y *La gran sultana*); otras tres presentarían el conflicto en torno al drama de una ciudad asediada (*La conquista de Jerusalén*, *El gallardo español* y *La Numancia*); dos se vertebrarían sobre un conflicto matrimonial (*La casa de los celos* y *El laberinto de amor*); tres se acercan notablemente al universo picaresco, con su correspondiente conflicto entre el individuo y su entorno (*La entretenida*, *El rufián dichoso* y *Pedro de Urdemalas*). Y otro tanto similar pasaría con los entremeses (García Aguilar, Gómez Canseco y Sáez 2016). Como se ve, este esquema justifica a la perfección la cercanía entre *El gallardo español*, *La conquista* y *Los tratos*, pues el motivo de la ciudad asediada, unido a los términos religiosos del conflicto presentado, explicaría sobradamente la conexión.

15. Sí, quizá, socio-política, al vincular los hechos de la España pasada con los de la España presente (García Aguilar, Gómez Canseco y Sáez 2016: 51). También hay conflicto entre el individuo y la identidad colectiva de la ciudad; e, incluso, entre dos visiones opuestas de ver el mundo: el pragmatismo racional del ejército romano frente al idealismo extremo, y a veces volcánico, de los numantinos (Rey Hazas 1992: 69-91). Pero ningún conflicto religioso.

equidad entre los candidatos en perjuicio de Cervantes para ver si, de esta forma, la estilometría era capaz por sí misma de trascender las semejanzas textuales entre las obras.

Y así se ha visto. Analizada *La conquista de Jerusalén* en el marco de un corpus, digámoslo así, poco favorable para la hipótesis cervantina, el resultado vuelve a ser el mismo: obras de Cervantes correctamente distribuidas y *La conquista* bien incluida entre sus filas (Figs. 4 y 5).

Si a esto añadimos la confirmación de las pruebas de clasificación (Figs. 6 y 7), atendiendo especialmente a los elevadísimos porcentajes de evaluación correcta, podremos concluir, ya con una mínima probabilidad de error, que *La conquista de Jerusalén* está escrita en un estilo y con un pulso netamente cervantinos. No es la interpretación aislada de los resultados, sino su visión en conjunto, lo que nos permite hablar, quizá no de pruebas, pero sí de unos niveles de probabilidad tan elevados que anulan, de facto, cualquier otra posible atribución. Y concluyamos diciendo que todos, absolutamente todos los escenarios observados a lo largo de este itinerario estilométrico, certifican y corroboran que la paternidad de la obra corresponde, efectivamente, a Miguel de Cervantes.

## Bibliografía citada

Antonucci, Fausta (2015). «La estructura dramática del teatro cervantino de la primera "época": una propuesta de análisis», *Cuadernos AISPI*. 5, pp. 131-146.

Arata, Stefano (1989). *Los manuscritos teatrales (siglos XVI y XVII) de la Biblioteca de Palacio*. Pisa: Giardino.

Arata, Stefano (1991). «Loyola y Cepeda: Dos dramaturgos del Siglo de Oro en la Biblioteca de Palacio», *Manuscrt. Cao* IV, pp. 3-15.

Arata, Stefano (1992). «La conquista de Jerusalén, Cervantes y la generación teatral de 1580», *Criticón*. 54, pp. 9-112.

Arata, Stefano (1997). «Notas sobre *La conquista de Jerusalén* y la transmisión manuscrita del primer teatro cervantino», *Edad de Oro*. 16, pp. 53-66.

Argamon, Shlomo; Anat R. Shimoni y Moshe Koppel (2002). «Automatically Categorizing Written Texts by Author Gender», *Literary and linguistic computing*. 17 (4), pp. 401-412.

Blasco, Javier (2016). «Avellaneda desde la estilometría», en Pedro Ruiz Pérez (ed.), *Cervantes: los viajes y los días*. Madrid: Sial Ediciones, pp. 97-116.

Blasco, Javier, Patricia Marín Cepeda y Cristina Ruiz Urbón (eds.) (2010). *Hos ego versiculos feci... Estudios de atribución y plagio*. Madrid: Iberoamericana - Vervuert.

Blasco, Javier y Cristina Ruiz Urbón (2009). «Evaluación y cuantificación de algunas técnicas de "atribución de autoría" en textos españoles», *Castilla: Estudios de literatura*. 0, pp. 27-47.

Brioso Santos, Héctor (2009). «A propósito de la historicidad de *La conquista de Jerusalén*: los cuatro milagros de la primera cruzada», *Anuario de Estudios Cervantinos*. 5, pp. 101-124.

Calvo Tello, José (2016). «Entendiendo Delta desde las Humanidades», *Caracteres. Estudios culturales y críticos de la esfera digital*. 5 (1), pp. 140-176.

Calvo Tello, José y Juan Cerezo Soler (2018). «La conquista de Jerusalén ¿de Cervantes? Análisis estilométrico sobre autoría en el teatro del Siglo de Oro español», *Digital Humanities Quaterly*. 12 (1).

Calvo Tello, José, Daniel Schlör, Ulrike Henny-Krahmer and Christof Schöch (2017). «Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels». Montreal: ADHO, August 8-11, pp. 181-183.

Eder, Maciej, Mike Kestemont y Jan Rybicki (2016). «Stylometry with R: A Package for Computational Text Analysis», *The R Journal.* 16 (1), pp. 1-15.

Fradejas Rueda, José Manuel (2016). «El análisis estilométrico aplicado a la literatura española: las novelas policiacas e históricas», *Caracteres*. 5 (2), pp. 196-245.

García Aguilar, Ignacio, Luis Gómez Canseco y Adrián J. Sáez (2016). *El teatro de Miguel de Cervantes*. Madrid: Visor Libros.

Hernández Lorenzo, Laura (2017). «Quantitative Syntactic Approaches to Spanish Poetry. A Preliminary Study on Fernando de Herrera's Poetic Works», en *The Educational Impact of DSE*. Roma: pp. 152-154.

Jannidis, Fotis, Steffen Pielström, Christof Schöch y Thorsten Vitt (2015). «Improving Burrows' Delta – an Empirical Evaluation of Text Distance Measures», en *Digital Humanities 2015 Conference Abstracts*. Sydney: ADHO, pp. 100-113.

Jockers, Matthew L. (2014). *Text Analysis with R for Students of Literature.* Springer.

Juola, Patrick (2015). «The Rowling Sase: A Proposed Standard Protocol for Authorship Attribution», *Digital Scholarship in the Humanities*. 30 (suppl. 1), pp. 100-113.

Kestemont, Mike (2011). «Een stylometrisch onderzoek naar Jan van Boendales auteurschap voor de Brabantse yeesten», en *Revue belge de philologie et d'histoire. Belgisch tijdschrift voor filologie en geschiedenis*. 89 (3-4), pp. 1019-1048.

Mosteller, Frederick y David L. Wallace (1963). «Inference in an Authorship Problem», *Journal of the American Statistical Association*. 58 (302), pp. 275-309.

Oakes, Michael (2009). «Corpus Linguistics and Stylometry», en *Corpus Linguistics: an International Handbook*, Anke Ludeling y Merja Kyto (eds.). Berlín: Mouton de Gruyter, pp. 1070-1090.

Rey Hazas, Antonio (1992). «Cervantes y Lope ante el personaje colectivo: *La Numancia* frente a *Fuenteovejuna*», en *Cervantes y el teatro. Cuadernos de Teatro Clásico*. 7, pp. 69-91.

Rißler-Pipka, Nanete (2016). «Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales», en *El otro Quijote. La continuación de Avellaneda y sus efectos*, Hanno Ehrlicher (ed.), mesa redonda. Augsburgo: Universität Augsburg, pp. 27-51.

Rojas Castro, Antonio (2017). «Luis de Góngora y la fábula mitológica del Siglo de Oro: clasificación de textos y análisis léxico con métodos informáticos», *Studia Aurea*. 10, pp. 111-142.

Rojo Alique, Pedro C. (1996-1998). «Notas acerca del catálogo de manuscritos de la Biblioteca del Palacio Real de Madrid», *Manuscrt. Cao*. VII, pp. 83-131.

Rosa, Javier de la y Juan Luis Suárez (2016). «The Life of *Lazarillo de Tormes* and of his Machine Learning Adversities. Non-Traditional Authorship Attribution Techniques in the Context of the Lazarillo», *Lemir*. 20, pp. 373-438.

Schöch, Christof (2013). «Fine-Tuning our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater», en *Digital Humanities 2013: Conference Abstracts*. Lincoln: UNL.

Shakespeare, William (2016). *The New Oxford Shakespeare: The Complete Works*. Modern Critical Edition, first edition, Gary Taylor, John Jowett, Terri Bourus y Gabriel Egan (eds.). Oxford: Oxford University Press.

Smith, Peter W. H. y W. Aldridge (2011). «Improving Authorship Attribution: Optimizing Burrows' Delta Method», *Journal of Quantitative Linguistics*. 18 (1), pp. 63-88.
Stamatatos, Efstathios (2009). «A Survey of Modern Authorship Attribution Methods», *Journal of the Association for Information Science and Technology*. 60 (3), pp. 538-556.

Recibido: 3 de octubre de 2017
Aceptado: 28 de agosto de 2018